INSTITUTE FOR DEFENSE ANALYSES

Real-Time Information Extraction from Big Data

Robert M. Rolfe, *Project Leader*

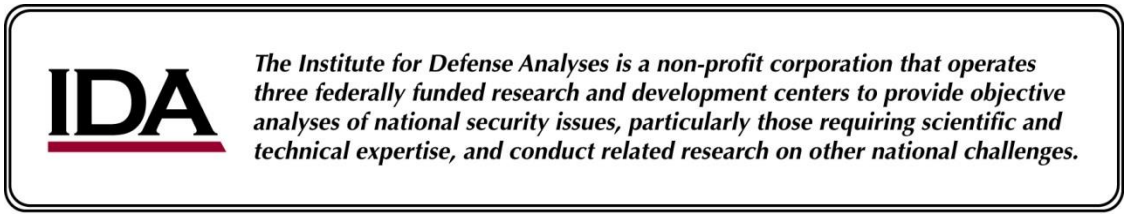
Jagdeep Shah

Francisco L. Loaiza-Lemos

October 2015

IDA Non-Standard Document
NS D-5618
Log: H 15-000986
Copy

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication

This work was conducted under IDA's independent research program (C5170). The views, opinions, and findings should not be construed as representing the official position of the Department of Defense.

Real-Time Information Extraction from Big Data

Jagdeep Shah
Robert M. Rolfe
Francisco L. Loaiza-Lemos

October 7, 2015

INSTITUTE FOR DEFENSE ANALYSES

Abstract

We are drowning under the 3 Vs (volume, velocity and variety) of big data. Real-time information extraction *from big data, at global as well as local levels, is critical for making rapid decisions in many important DoD scenarios. However, such real-time information from big data is enormously complex and extremely challenging. We argue that* data movement *is of crucial importance in big data analytics, and show that recently developed photonic communication technologies has the potential to extract real-time information, and improve performance/Watt by a hundredfold with corresponding cost savings at the local, rack-size level, provided the entire system, consisting of communications hardware and network, computer hardware, algorithms, software and architecture, is* holistically optimized. *We provide examples of some applications of interest to DoD for which such holistic system optimization could be a game changer.*

Summary

The Department of Defense (DoD) needs to be able to make rapid decisions based on (a) the enormous amounts of data streaming in at blindingly fast rates from its vast collection of sources and sensors distributed over the globe, and (b) on the correlations of the streaming data with historical data stored locally and globally. Such real-time decision support is a highly challenging problem requiring technical and conceptual advances at multiple levels. Optimization of these multiple aspects for real-time decision support is critically important for DoD in areas such as real-time command and control (C2), command, control, communications, intelligence, surveillance, and reconnaissance (C4ISR), planning and analyses for C2 decision support, cognitive big data, cloud robotics, and advanced autonomous systems.

The world is drowning under the three Vs of big data: volume, velocity and variety. Real-time information extraction from such big data is now recognized as the future foundation for all the applications that are expected to generate quick and timely decision support to DoD decision makers. Real-time information extraction constitutes a daunting challenge and is beyond the capabilities of current systems, mainly because they are optimized for spatially localized data access patterns, whereas most big data applications require access to small bytes of non-local data randomly distributed over a vast memory and geographic space. Therefore, data communication is critically important in big data analyses at a single geographic location, as well as for geographically distributed locations. Furthermore, the type of data communications technologies used for distributed geographic data centers are fundamentally different from those used within nodes of localized data centers. Optimizing big data analytics at local and geographically distributed levels very likely will require different communications technologies, algorithms, architecture, and software because real-time information extraction from big data is inherently a multi-dimensional problem. But both levels must be addressed in order to adequately handle the richly variegated scenarios important to DoD.

Recent developments in photonic technology can provide the critically important communications hardware and networks underlying efficient big data analytics, which is often performed by Graph Processors capable of consuming graphs with large numbers of vertices and connecting edges. Optimization of individual aspects of Graph Processors, such as algorithms and architecture, have been considered for existing hardware and computing systems. We argue that it is essential to optimize the entire Graph Processor system holistically; whether geographically localized or geographically distributed, it is critical to co-optimize the photonic technologies for the essential high-bandwidth, high-capacity, energy-efficient communications with computer hardware, algorithms, architecture, software, and network control system. Such a holistically optimized Graph Processor could potentially provide the required real-time information extraction capability for big data problems of interest to DoD with more than two orders of magnitude

improvement in performance/Watt for geographically localized nodes, and the corresponding cost benefits. We discuss some likely applications of interest to DoD and conclude with a discussion of who would be interested in such real-time information extraction capabilities and what are the next steps in such analyses of real-time decision support for applications of interest.

1. Introduction

Enormous amounts of data are being generated by a large number of sensors and devices (Internet of Things: IoT), and this data is being gathered via various collection techniques. But it is not just the volume of data that is increasing, the rate, i.e., the velocity at which data is being collected, is also increasing. Furthermore, these data sets are becoming more complex and diverse and therefore the types of data are also increasing. The world is drowning under these three Vs (Volume, Velocity and Variety) of big data.

The problem is highly complex and, on a worldwide level, it requires optimization of global networks, algorithms, and architecture for parallel processing over globally distributed computing systems. On a local level, it requires optimization of a local computing system for rapid, real-time information extraction for decision support. Many techniques and efficient systems have been developed to *search* for a given term or phrase within a large database. But many problems of interest to DoD, as well as to other U.S. Agencies and the commercial world, are not simple search problems; i.e., problems that identify whether or how many times a particular term or phrase is encountered in the data set. We are interested more and more in *extracting information that is implicit in the big data* and not just querying a database for a given term. Furthermore, *real-time information extraction* is of enormous value, and even essential, to enable rapid decisions in time-critical situations.

Current computer systems are woefully inadequate for such real-time information extraction from big data. The reason for this, in a nutshell, has to do with the difference between the data access patterns used for problems addressed by current computer systems and those required in the emerging big data problems. We will examine these differences in Section 2.

The U.S. Government has recognized the big data problem, and a Presidential initiative was launched in 2012 to address these issues. A number of these activities have targeted the algorithmic and architectural issues related to computer systems handling big data problems. While this is important work and is clearly needed, we posit that *efficient underlying communications technologies* are essential for the optimal resolution of big data problems. In Section 3, we surmise that newly developed photonics technologies can be a game changer by providing high bandwidth, high-capacity, energy-efficient communications and communications networks between a large cluster of computer nodes with the large memory required to store high data volume, and they can form the foundational technology for tackling big data problems effectively.

While photonics technologies can provide the essential underlying hardware and network technology, it is not enough by itself to address the big data challenge. We contend in Section 4 that an *optimization of the entire system* is required for ideal real-time information extraction from big data. This includes incorporating the critical underlying photonics communications at both the local and global levels, as well as computer processing and storage hardware, computer network control logic, algorithms, architecture and applications software.

Many of the big data problems can be formulated in terms of Graph Problems,[1,2] in which data is represented by the vertices and the connecting edges of a graph. Therefore, a holistically optimized Graph Processor (HOGP)[3] for real-time information extraction from big data would be of enormous value; we discuss important aspects and ingredients of a HOGP in this document. On a local level, the fundamental issue is to provide energy-efficient data movement for small bytes of data randomly located across a large (Petabytes or greater) memory space. On a global level, the fundamental issue is to map the algorithms onto systems hardware to achieve parallelization of processing to dramatically improve system throughput. This was described in a 1990 paper[4] by Professor Leslie Valiant, and we quote him here:

> The success of the von Neumann model of sequential computation is attributable to the fact that it is an efficient bridge between software and hardware: high-level languages can be efficiently compiled onto this model; yet it can be efficiently implemented in hardware. The author argues that an analogous bridge between software and hardware is required for parallel computation if that is to become as widely used. This article introduces the bulk-synchronous parallel (BSP) model as a candidate for this role, and gives results quantifying its efficiency both in implementing high-level language features and algorithms, as well as in being implemented in hardware.

A processing system capable of real-time information extraction can be a game changer for many problems of interest to DoD. These include: real-time command and control (C2) decision support systems for cyber warfare, command, control, communications, intelligence, surveillance, and reconnaissance (C4ISR), planning and analyses for C2 decision support, and cognitive big data and cloud robotics and autonomy. We discuss in Section 5 how these applications may benefit from real-time information extraction and

---

1 https://en.wikipedia.org/wiki/Graph_database

2 http://hama.apache.org/

3 https://en.wikipedia.org/wiki/Graph_theory

4 Communications of the ACM, August 1990 vol. 33, No 8 page 103.

which agencies might be interested in such capabilities. In Section 6 we discuss other federal applications.

We outline in Section 6 the steps that can be taken to make the arguments and discussion presented in this document more robust, and present a brief summary in Section 7.

2. Data Access Patterns for Current and Big Data Systems

Many current solution architectures rely on accessing data resident in related and close storage locations, e.g., nearby locations in a memory. Current computer systems are designed to take advantage of such *spatial locality of data*. Thus a cache line, consisting of 128 Bytes of data, is downloaded from a remote memory location and stored in cache memory on the processor chip, with the expectation that the processor very likely will require the data so stored in the cache for the next several operations and that the likelihood of a "cache miss", i.e., not finding the data in the cache, will be small. For this class of problems, the computation is thus performed efficiently while minimizing the latency in execution of algorithms while possibly increasing data communication loads.

In contrast, big data problems require repeated access to a few bytes of data that are randomly located across a vast memory space,[5] e.g., across either a rack, a warehouse-size data center, or across many global locations. So downloading a cache line of data at a time and storing it in a local cache will not improve the overall performance. Also, the memory space is vast because of the large volume of big data, and the few bytes of data required for successive calculations can be randomly located across this vast memory space. Therefore, efficient solution of many big data problems require efficient communication of such data to the processors in a computing node or processors in globally distributed computing nodes. Furthermore, many of the problems are not completely parallelizable, so intermediate results store in a computing node also may be required at another computing node to continue the analysis, which necessitates further data movement between computing nodes. Thus, while some judicious distributed processing would be helpful, efficient data movement requirements cannot be totally eliminated.

Much progress has been made on electrical communications in computer clusters, but the current solutions cannot provide the bandwidth, the capacity, and energy efficiency necessary for managing the enormous data movement necessary for many big data problems, either at a local or at a global level. In recent years, significant progress has made on Silicon Photonics through the Defense Advanced Research Projects Agency (DARPA) and other support. Silicon photonics can provide high-bandwidth, high-capacity, energy-efficient communications in small form factors and can be fabricated in processes fully

[5] Memory address space and spatial distribution of memory stores create slowdowns in algorithmic data processing. When we are not focused on localized memory space, we will explicitly communicate to the reader the nature of the architecture that is not localized to centralized large processing nodes.

compatible with silicon electronics. Large-scale manufacturing of silicon photonics is being addressed by the newly formed Integrated Photonics Manufacturing Innovation Institute (IP-MII). In the next section we provide a brief overview of silicon photonics technologies and how a HOGP with underlying silicon photonics technologies, performed by IBM under the DARPA Photonically Optimized Embedded Microprocessors (POEM) Program, could potentially deliver efficient big data processing.

3. Photonic Technology for Data Access Patterns of Big Data

DARPA's POEM Program supported much of the development of Silicon Photonics technology and also investigated in detail the design and performance of a holistically optimized petascale, rack-size computer system for addressing the big data problem using underlying silicon photonic technologies; i.e., a photonically optimized Graph Processor. This design and analysis was performed by IBM's TOPS program under DARPA's POEM Program.

Photonic communications have revolutionized the world of communications beginning with global or long-distance and then moving to metropolitan and local networks, and now between racks and units within a rack. These photonic technologies primarily use efficient III-V semiconductor devices as transmitters, receivers, and mux/demux and optical fibers as communications media. Communications on a chip (such as a processor chip), between chips (such as processor and memory chips), on a board, between boards, and within a petascale rack (Petabytes of memory, Petaflops processing power within a rack) requires very different photonic technologies.

Silicon photonic technologies are well-suited for seamless communications from the on-chip source of data to other on-chip locations and from on-chip to other chips, and on a processor board. Although Silicon is not the most efficient photonic material, it brings many advantages to the table, among them compatibility with silicon electronics and the ability to leverage silicon electronics manufacturing infrastructure. DARPA's POEM program has shown that a complete suite of high-performance silicon photonic devices can be fabricated in industrial foundries with fabrication processes fully compatible with high-performance electronics. Compatibility with electronic processing allows co-design of photonic and electronic devices on a single chip, leading to compact functional units with high-energy efficiency.

This silicon photonic technology is fully capable of providing the high bandwidth, high capacity, energy-efficient communications necessary for the data access patterns required for big data analytics—small bytes of data, randomly distributed over a large (Petabytes of larger) memory. IBM has analyzed a rack-size, petascale cluster of 64 nodes, each with ~ 16 TFLOPS of processing power and ~ 16 TB of memory using silicon photonic transceiver, switch, and amplifier technologies (see Figure 1). The nodes are connected with efficient and controllable optical switch planes. IBM has optimized the performance

of such a Graph Processor by co-optimizing photonic technology, network, integrated circuit, and network architecture and system control. IBM has compared the performance of such a rack-size, petascale, photonically optimized Graph Processor with that of Blue Gene Q, one of the two highest-rated Top 500 computer systems in the world and which has much larger processing power. They find that for the generally used Graph 500 benchmark GTEPS and GTEPS/kW, this photonically optimized rack-size petascale cluster outperforms Blue Gene Q, occupying a large floor by a factor of 4 in GTEPS (GigaTraversedEdgesPerSecond) and a factor of nearly 200 in GTEPS/kW. Furthermore, latency is a most important parameter for real-time information extraction, and latency for such a cluster would be an order of magnitude shorter than the Blue Gene Q. These benefits primarily stem from the efficient data movement provided by the photonic technologies for the data access patterns relevant to the Graph 500 algorithm benchmarks. Such novel approaches have the potential for real-time information extraction critical for many defense applications, with considerable savings in power consumption and hence cost.

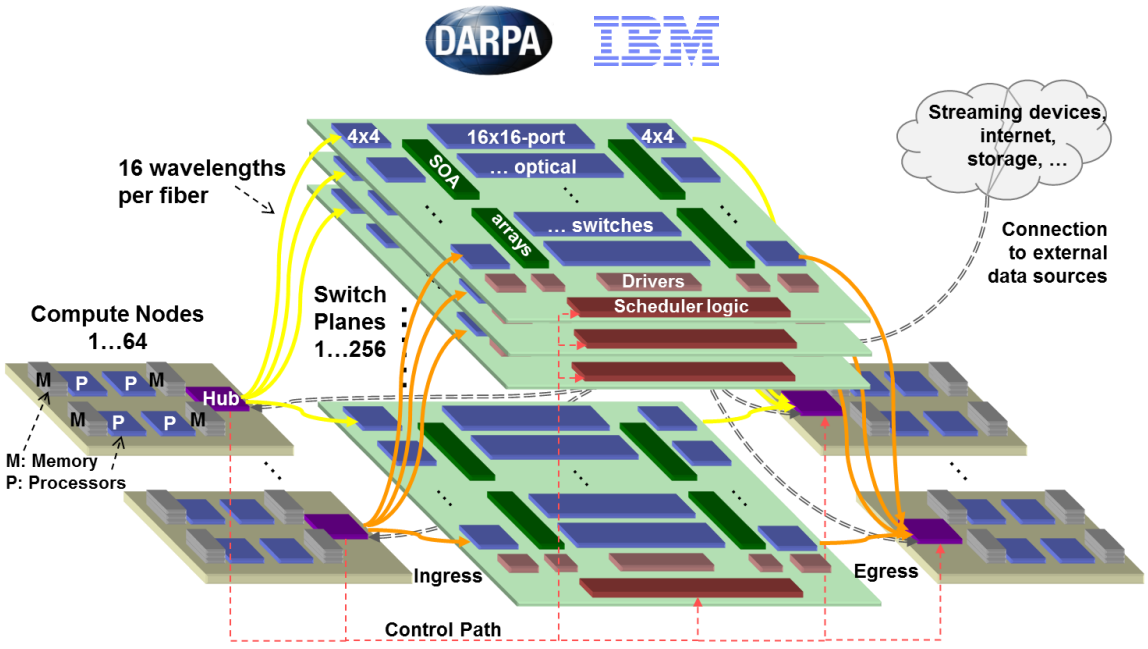


Figure 1. IBM's Design for a Rack-size, Petascale Cluster for Real-time Information Extraction from Big Data[6]

[6] IBM's TOPS contract under DARPA's POEM Program: L. Schares et al., "A Throughput-Optimized Optical Network for Data-Intensive Computing," IEEE Micro, No.5-6, (2014).

4. Holistic Optimization of Big Data Systems for Real-Time Information Extraction

It is important to emphasize that while photonic technology to enable efficient movement of small bytes of data randomly located across a large memory space was the essential underlying technology, other system optimization also played a key role in achieving the improved performance cited above. For example, the electronic control of data movement must be optimized and each node must be designed to minimize the inter-node data traffic for maximum energy efficiency and minimum latency.

The general point to emphasize is that the entire computing system must be *holistically optimized by co-optimizing all components of the entire system.* These components include: photonic hardware, photonic network, electronics for data control and processing, system architecture, algorithms, and software for big data applications. Optimal allocation of various system functions between hardware, algorithms, architecture, and software is crucial in a complex system design. Until now, big data analytics has addressed the algorithm, architecture, and software aspects using the existing communication technologies and networks, i.e., without the benefit of new communications technologies. Our novel approach recognizes and emphasizes for the first time that the entire system, as defined above, must be holistically optimized by considering newly developed communications technologies and computer system design in a cohesive manner.

This is an enormous amount of effort, and we are just scratching the surface here. But the photonic technology is ready for transition into a real system and holistic system optimization efforts are worthwhile and will have huge payoffs for many applications, including some critically important for DoD.

A HOGP capable of real-time information extraction would address many different types of problems. Anomaly detection is important in many different situations. For example, in cyber security one would constantly monitor the traffic patterns into and out of a set of servers. Historical data are stored over long periods of times and current traffic patterns need to be constantly compared to historic patterns to detect anomalies. Time is of the essence in detecting any anomalies in order to respond rapidly, for example, to a cyber-attack to protect the computer system. This is a big data problem because historic data can occupy very large memory space (volume of big data) and rapidly incoming traffic needs to be compared to the historic data.

Collection of images and other types of information by sensors is growing by leaps and bounds. Not only is the rate of collection of data increasing because of more advanced sensors, but the number of sensors of different types is increasing and real-time fusion of information and extraction of information from multiple sensor types is extremely valuable.

Information of this type is currently extracted by highly skilled human experts based on their intuition and vast knowledge. We do not have, and cannot produce enough experts to fill our requirements for rapid decisions, and the current computer systems are incapable of extracting information from big data in a timely manner. We must find a way to develop systems that leverage HOGPs and can thus extract real-time information from big data.

We note that a National Strategic Computing Initiative (NSCI) was launched recently by the President. The primary focus of NSCI is to demonstrate a computing system capable of Linpack Exaflops benchmark[7] performance with 20 MW power consumption, and with secondary focus on other issues. We can leverage the advances in power efficiency achieved in NSCI but it is important to emphasize that our focus is on real-time information extraction for rapid decision support, and that it involves optimization of the entire system, including communications, computing hardware and networks, architecture, algorithms, and software, and not just optimization for achieving Linpack Exaflops/MW targets.

5. DoD Applications for Real-time Decision Support

Real-time decision support is of enormous value to DoD. Big data analytics processes operate on the captured data with current latencies of minutes to days depending on the scope of the analytics. As observed in the news media, current cyber events may not be recognized for many months before assessments are obtained. The consequence is clear: cyberattacks may slip through and remain undetected for many months. Our focus is how to reduce these latencies to seconds or less, rather than minutes or days, i.e., real-time information extraction from big data for applications of interest to DoD. Real-time information from big data is an enormously complex with multiple challenges:

- Anticipate future adversarial and indigenous population activity and action
- Synchronize actions across air, space, and cyberspace
- Build trusted autonomous systems to enable machine-aided decision support
- Deliver agile C2 capabilities for future dynamic conflicts
- Provide rapid, continuous assessments of complex, cascading effects.

The challenging problem of real-time decision support for C2 has global, as well as local, aspects, and both must be analyzed for optimum solutions:

1. **Global Aspect:** Data flows in real time at very high rates from multiple sensors and joint bases through the global networks and needs to be captured and analyzed in multiple parallel processing centers. The data rate is so large that not all data can be captured, much less analyzed in real time. Furthermore, these streaming data may also need to be compared with historically stored data at the

[7] The LINPACK Benchmarks are a measure of a system's floating point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering. See https://en.wikipedia.org/wiki/LINPACK_benchmarks.

parallel processing centers. One problem we will address is the holistic optimization of this distributed parallel processing system. The holistic optimization will consider all aspects of the system, including computer algorithms, deployed architecture, and software, as well as communications hardware and networks.

2. **Local Aspect:** Equally important is real-time information extraction from data streaming into a data center and from its comparison with historical data stored at this data center. For data collected globally, as in Figure 1 above, some pre-analysis is performed at the distributed processing centers and selected pre-analyzed data is transmitted to a local data center and stored there as historical data for further processing. Alternatively, one may be interested in real-time information extraction of data streaming in at one local data center. A local data center may be a single rack with a high-performance petascale processor or hundreds of racks of processors distributed over a floor or a data warehouse. Once the data of interest is at a local data center, the problem is geospatially bound, but holistic optimization of the processing system must include consideration of the same components, i.e., computer algorithms, data center architecture, software, hardware for communications within the data center, and communication network. Photonic technology is an essential underlying technology for this case as well and can bring large performance and cost benefits as we have discussed above.

We give here a high-level overview of three applications of interest to DoD. A more detailed discussion of these applications is provided in Appendix A.

A. Cyberspace Command and Control Decision Support

Command and control decision support is about rapidly having the right information, at the right place, to make the right decision. This is a core function for Air Force and other services. There is an increasing need for real-time information extraction from big data in order to make C2 decisions more efficiently and effectively. This enormously complex problem is made even more challenging by the necessity of integrating the C2 capabilities across the air, space, and cyber domains, each with dramatically different characteristics of speed, time, and distance.

Many C2 decisions depend on big data analytics and its ability to extract information rapidly. Real-time C2 decision support for cyber defense can benefit from computing capabilities that are beginning to achieve the required technical maturity. At the moment rapid decision making is not possible because of the limitations of the communications hardware and networks as well as current processing systems. Data are streaming in at an increasingly rapid rate and in increasing volume from joint bases and sensors located around the globe.

One example of the inadequacy of current systems is the Joint Regional security stacks (JRSS) that each operate in near real-time at 40 GB/second. While the data flows in real-time from joint bases, the MPLS network is quantized in 100 GB/second data pipes, and further broken down by the MPLS Label Edge Routers to real-time manageable slices of 40 GB/sec that enter specific JRSS stacks. The data that can be processed in real time is processed and at the end of the stack that data reenters the network and sent on its way to meet the intended purpose unless the real-time decision inhibits further transmission. In parallel, the real-time data is captured along with real-time logs from the stack processes and this data is stored for subsequent data analytics or pre-processed and sent to a local data center for further analysis.

Big data analytics currently processes the captured data with latencies of minutes to days depending on the scope of the analytics, with the result that cyberattacks may slip through and remain undetected for many months. The latency in detecting cyber events is created by combination of processes that are autonomous but cannot keep up with real-time data flows and processes that require human–machine interaction and big data analytics. Some cyber offense processes may require real-time construction and analyses of network graph models that help visualize dependencies that may require uncertainty quantification and analyses with a variety of inference engines on knowledge bases that are built and expanded continuously. The computation system to meet the C2 needs of the networks is vast and dispersed and often laden with incompleteness and/or uncertainty. Figure A-1 in Appendix A illustrates an example of the concept of operation (CONOPS) for deployment of the resources required to obtain situational awareness in cyberspace and thus to support adequate command and control capabilities.

B. C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance) Net Assessment and Analyses for C2 Decision Support

With the robust C4ISR capability of our networks, there is a need for decision support processes that can use the information and convert it to battlefield knowledge. In future warfare, we need to develop C4ISR capabilities to rapidly answer complex questions for the warfighter. The big data analytics must address the totality of forces (friendly, enemy, neutrals etc.), as well as systems and populations including the specific country "net assessments," intelligence preparation of the battlefield, and on-going intelligence-driven operations. In a battlefield, thousands of nodes share a common view through a system of localized caches and a centralized core database in the theatre. It would be desirable to have real-time computation nodes that could process nodal databases in real time to construct and employ knowledge bases.

One of the keys to success is to employ big data analytics across hundreds of data sources to answer warfighter questions and to discover how events, people, and places are related. The key to supporting this warfare is to train with the full C4ISR capability at hand as you fight.

The important consideration for big data analytics in these situations is not the large number of dispersed nodes (thousands) and networks. It is rather the need to perform cognitive computing at a few nodes with huge numbers of highly interconnected nodal processors with cross-connecting backplane bandwidths exceeding many TB/sec. In this case nodal processors requiring two orders of magnitude less energy and occupying at least one order of magnitude less space, possible with a HOGP, would be real game changers by enabling automated cognitive decision support. The goal is to provide each leader access to these cognitive decision aids. To do this, we must estimate the data flows and processing requirements that could be achieved with the proper combination of standard processors, data center techniques of interconnect, and a specialized array of capable HOGPs for knowledge base operations that include querying, reasoning, and cognitive analyses.

C. Cognitive Big Data and Cloud Robotics and Autonomy

Integrated cognitive big data processing is the holy grail of big data analytics. The ultimate goal of this use case to simultaneously support human and machine together to achieve any mission. We need big data analytics and integrated cognition to provide teams of human and cognitive machines with the knowledge they need to achieve mission success.

6. Other Federal Applications for Real-time Decision Support

A. Organization Knowledge Management and Retention

DoD, as well as other parts of the Federal Government, will experience a massive loss of corporate knowledge as personnel within the current work force who are in the over-60 age group begin to retire. Recent developments in machine learning using deep neural networks[8] show that it is possible to configure systems that by examining relevant patterns can extract the embedded knowledge of a particular area of expertise and essentially teach themselves how to perform specific tasks. Obviously, the volume of subject matter expertise is enormous, and its complexity and variety qualifies it as a big data problem. Here too, the geospatial spread of the knowledge collected variously in internal reports, user manuals, and graph databases, would benefit from implementations based on HOGPs than can achieve Exaflop/MW performance benchmarks.

B. Physical Security

The DoD as well as a number of Federal Agencies own and man large numbers of installations—ranging from radars to power stations for data centers. Most of these installations have very little physical protection—a chain link or a chicken wire fence. This is likely to change, and the number of sensors that will be feeding data in real time will likely be massive.

[8] Giraffe – Using Deep Reinforcement Learning to Play Chess.pdf.

C. Border Security

The extensive borders that the United States has with Mexico and Canada may represent a tremendous weakness in our overall physical security if we are ever confronted with immigration levels similar to those Europe is facing right now. Attempts to seed large stretches of the border with seismic and other kinds of sensors could likely benefit from more powerful data processing—along the lines of what is proposed in the this report.

D. Intelligence Collection and Exploitation Applications

Big data associated with collection and exploitation of unclassified sources for intelligence, e.g., newspapers, TV, political blogs, and technical journals. Although this problem also requires advanced semantic techniques and Natural Language Processing (NLP), the sheer volume and variety poses a huge challenge in terms of correlation.

7. Next Steps

These are very critical applications and needs, but no existing system can provide the required capabilities for real-time information extraction for rapid decision-making and response. Expert human analysts are too few in number and are being overwhelmed by the 3Vs of big data. The only solution is to develop computer systems capable of providing such capabilities more quickly and efficiently than our rivals and/or potential enemies. A HOGP with underlying essential photonics technology providing the high-bandwidth, high-capacity, and high-efficiency communications essential for addressing the data access patterns of big data analytics has the potential for providing the required critical capabilities. But much more work remains to be done. The next steps in the process are as follows:

- Explore further applications as required to meet the sponsoring agency's needs for parallel computing,
- Provide example models of BSP holistic optimization to support analyses of specific applications,
- Assess the application of BSP model to systems of systems for specific applications by:
  - Identifying the key elements of such a holistic optimized system,
  - Determining the requirements for specialized graph processors,
  - Defining a conceptual design of a prototype that can help refine the benefits from utilization of an HOGP,
- Determine the best research and development path to meet critical needs.

8. Conclusions

We posit that data communications will be a big bottleneck for big data analyses, and that newly developed photon-based inter-processor communications technology could help

provide the required enhanced system communications. But improving communication efficiency is just one aspect of the solution. It is essential to optimize the entire processing system, including the architecture, employment of advanced photon technology, enhanced global communication networking as required, computer hardware, and software algorithms. Such a holistically optimized system could deliver a real-time information extraction capability with large improvements in the processing per Watt of energy and the corresponding cost savings. This could be a game changer for many applications of interest to DoD and federal agencies.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
00-10-15	Non-Standard	

4. TITLE AND SUBTITLE
Real-Time Information Extraction from Big Data

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBERS

6. AUTHOR(S)
Jagdeep Shah
Robert M. Rolfe
Francisco L. Loaiza-Lemos

5d. PROJECT NUMBER
C5170

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
NS D-5618
H 15-000986

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive, Alexandria

10. SPONSOR'S / MONITOR'S ACRONYM
IDA

11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Project Leader: Robert M. Rolfe

14. ABSTRACT

We are drowning under the 3 V's (volume, velocity and variety) of big data. Real-time information extraction from big data, at global as well as local levels, is critical for making rapid decisions in many important DoD scenarios. However, such real-time information from big data is enormously complex and extremely challenging. We argue that data movement is of crucial importance in big data analytics, and show that recently developed photonic communication technologies has the potential to extract real-time information, and improve performance/Watt by a hundredfold with corresponding cost savings at the local, rack-size level, provided the entire system, consisting of communications hardware and network, computer hardware, algorithms, software and architecture, is holistically optimized. We provide examples of some applications of interest to DoD for which such holistic system optimization could be a game changer.

15. SUBJECT TERMS
Big data, big data analytics, real-time information extraction, holistically optimized graph processor (HOGP), photonics, Command and Control (C2), Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR), open source intelligence, corporate knowledge retention, cyber intrusion detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include Area Code)
Unclassified	Unclassified	Unclassified	Unlimited	14	